Yixue Tongjixue

高等医药院校教材

医学统计学

（第2版）

⊙陈启光 主编

东南大学出版社

高等医药院校教材

医学统计学

（第2版）

主　编　陈启光

编　者　（按章节顺序）

陈　峰　（南京医科大学）

顾海雁　（南通大学）

于　浩　（南京医科大学）

娄冬华　（南京医科大学）

杨永生　（苏州大学）

闵　捷　（东南大学）

刘　沛　（东南大学）

陈全良　（南京中医药大学）

唐　尧　（扬州大学）

孙　峰　（扬州大学）

黄水平　（徐州医学院）

刘玉秀　（南京军区南京总医院）

李君荣　（江苏大学）

陈炳为　（东南大学）

陈启光　（东南大学）

东南大学出版社

南　京

内 容 提 要

全书内容分绪论、统计资料的整理与描述、统计表与统计图、研究设计基础、统计推断基础、两组资料及多组资料均数比较、两个率或多个率的比较、非参数统计方法、相关与回归、临床试验中的统计学应用基础、诊断试验评价、随访资料统计分析以及统计方法的综合运用等，书末附有15个统计用表，每章后附有小结和复习思考题。

本书适用于医学院校除预防医学专业外各专业各层次学生以及成人继续教育学生，也可以作为参考书供临床医师使用。

图书在版编目(CIP)数据

医学统计学/陈启光主编．—2版．—南京：东南大学出版社，2007.3

ISBN 978-7-5641-0641-6

Ⅰ.医… Ⅱ.陈… Ⅲ.医学统计-医学院校-教材 Ⅳ.R195.1

中国版本图书馆CIP数据核字(2007)第006373号

东南大学出版社出版发行

(南京四牌楼2号 邮编210096)

出版人：江 汉

江苏省新华书店经销 常熟市华顺印刷有限公司印刷

开本：787mm×1092mm 1/16 印张：13.25 字数：347千字

2007年3月第2版 2007年3月第6次印刷

ISBN 978-7-5641-0641-6/R·72

印数：18001～23000 定价：19.00元

(凡因印装质量问题，可直接向读者服务部调换。电话：025-83792328)

再版前言

《医学统计学》自2002年出版以来已有五年多了。该书通过五年的教学实践，得到了教师和学生的鼓励和支持。他们普遍反映该书内容简洁严谨、重点突出，思路清晰，实用性强，适用于医学专业本、专科及成人教育的医学统计学课程教学。同时他们也对该书提出了很多宝贵意见。

原书的编委们通过教学实践并接受广大读者对该书所提出的意见，经过认真讨论，认为有必要对该书作修订。我们希望修订后的教材除了能继续保持原教材的特点外，在内容上还能反映医学统计学的学科发展，并对学生今后临床工作有实用价值。为此，我们将原书的第11章内容更换为“临床试验中的统计学应用基础”；将第13章“医学随访资料的统计分析方法”篇幅适当压缩，使重点更突出；为了便于读者掌握各章的内容，在每章后增加了各章小结。

本书在修订的过程中得到东南大学出版社和张慧编辑的大力支持，我们表示衷心感谢！

我们敬请广大师生和读者对本书不足之处提出宝贵意见。

编　者

2007年1月

目　录

1 绪论

1.1 引言

客观世界总是处于永恒的变化之中，只有从变化中去认识世界，才能对它有深刻的了解。事物的变化就其性质来说，有量变与质变之分，在质和量的密切联系中不断发展；就其变化的现象来说，有必然和偶然之分，且往往是偶然性（不确定性）掩盖了必然性，妨碍了我们对客观规律的认识。统计学的根本任务，在于揭露隐藏在偶然现象背后的必然性，是认识世界的重要手段。具体地说，统计学是研究数据的搜集、整理与分析的科学，面对不确定数据作出科学的推断或预测，直至为采取一定的决策和行动提出依据和建议。

人类实践是统计学产生的源泉，人类认识又是统计学发展的动力。统计学起源于17世纪中叶，最初的统计是一种计数活动，意指事实与数据，称为古典统计学。在西方，统计学一词源于state，意指对各个国家国情的叙述，内容无非是人口、财富和军事等，其研究方法主要采用形式逻辑的比较法和文字记述。18世纪后叶，开始重视数字资料和图表描述，标志着近代统计学的开始，其研究方法主要是建立在大样本上的大量观察法。其间，误差理论和大数法则得到了发展。然而，大量观察法并非适用于所有情况，例如，武器的试验，某些产品质量的检查等不容许也不可能进行大量的实验观察，其局限性和不足在应用中不断暴露。直到1908年，英国统计学家戈赛特（WS. Gosset，1876—1937）在Biometrika杂志上以笔名student发表了 t 分布，开始了小样本的研究，从而使统计学由“描述统计”向“推断统计”发展，开创了现代统计学的新纪元。20世纪50年代，电子计算机技术的发展和普及，促进了统计方法的应用和发展。当今信息社会，对有效地搜集数据，进行精确分析和可靠推断，作出科学决策，有着广泛的需求，统计学原理和方法几乎应用到自然科学和社会科学的各个领域，产生了许多应用性分支，如经济统计学、工业统计学、农业统计学、生物统计学等。

医学统计学是以医学理论为指导，借助统计学的原理和方法研究医学现象中数据的搜集、整理、分析和推断的一门应用性学科。

1.1.1 医学统计学的主要内容

医学研究和统计学的关系日益密切，可以说，几乎没有一个医学科研项目用不上统计思维与方法；同时，几乎所有的统计学原理与方法均可在应用中找到直接或间接的用途。根据目前医学研究的现状，本书着重介绍以下主要内容：

1）医学研究统计设计　进行医学科研设计，除应用必要的专业知识外，必须应用医学统计设计的基本原理，对实验的每个环节进行周密设计，目的在于创造一致的对比条件，有效地控制试验误差，以较少的人力、物力和时间取得较好的效果。详见第4章。

2）分布理论　是统计学的基础理论，主要用于探讨疾病的统计分布规律，为选择相应的

统计分析方法(如假设检验、统计建模、质量控制、疾病监测方法等)提供依据,是制订临床参考值范围,研究疾病等在空间上、时间上或人群中的分布规律的重要手段。详见第5章。

3) 统计描述　对原始资料进行一般性的描述,以期得到初步的了解和直观印象(如平均水平、离散程度、分布形状等),可用文字或统计图表表示。详见第2、3章。

4) 参数估计与假设检验　在大多数医学科研中需要对研究对象的全体(称之为总体)的某些参数(如均数、率、参考值范围等)作出适当的估计。参数估计是推断统计学的重要组成部分,假设检验是对资料是否来自具有某种属性的总体进行检验,常用于新药鉴定、病因分析、理化检验方法和技术水平的考核等,是推断统计学的重要组成部分,包括 t 检验、方差分析、χ^2 检验、秩和检验等。详见第6、7、8、10章。

5) 相关与回归　主要研究两变量之间的关系,常用于病因学研究、发育或生理功能评价,以及各种预测、趋势分析等,包括线性相关、直线回归、曲线拟合。详见第9章。

6) 新药临床试验与诊断试验　为了确定试验用药的安全性和有效性,我国已经颁布了药品临床试验管理规范(good clinical practice,GCP)。统计学作为保证新药临床研究科学性的重要手段也相应制订了指导原则。在第11章中简要介绍新药临床试验中的统计分析方法。此外,对机体的体液、细胞等进行化验是临床诊断的重要措施,诊断试验的科学评价是正确认识某诊断试验在临床上的应用价值的重要方法。详见第12章。

7) 随访资料的生存分析　"存活"时间的长短是临床随访研究工作者关心的主要指标。生存分析主要用于分析随访对象的生存规律及影响生存期长短的危险因素,并对截尾数据所提供的关于生存时间的不完全信息进行有效的分析,包括生存率的估计、Logrank 检验等。详见第13章。

8) 综合应用　学习统计学的目的是为了正确应用。本书对各种统计分析方法进行了系统的讲解,但在介绍各种统计分析方法时都是单一的,即一种方法用于解决一类问题。本书最后一章从实际出发,就综合应用各种方法解决具体的实际问题的分析策略进行了讲解,目的是为了帮助提高解决实际问题的能力。

1.1.2　学习医学统计学的目的与要求

学习统计学并非要使人们成为统计专业人员,其目的在于使大家具备新的推理思维,学会从不确定性和概率的角度去考虑问题;学会结合专业问题合理设计试验,通过精细的试验观察获得可靠、准确的资料;学会正确运用统计方法充分挖掘资料中隐含的信息,并能恰如其分地作出理性概括,写成具有一定学术水平的研究报告或科学论文,提高自身的科研素养。

为此,医药卫生各专业的学生必须学习医学统计学。

1.1.3　如何学好医学统计学

统计学的思维是用变异与不确定性、机遇与概率的观点去考虑问题,在相同的基础上去比较、分析,依据概率用逻辑推理去作结论,属归纳推理型思维。这在一定程度上与人们在其他学科学习和日常生活中养成的确定性的、偏于演绎推理型的思维方法有所不同,初学统计应注意这一点。

统计离不开数字,每个数字都有其实际意义。医学研究中,研究者收集到的数据从表面上看似乎杂乱无章,其间却可能隐含着内在的规律。因此不要厌烦数字,应重视原始数据的完整性和准确性,对数据处理持严肃、认真、实事求是的科学态度,反对伪造和篡改统计数据。

统计亦离不开公式和计算。统计学中的公式都是由实际问题引申出来的，一般都有其实际意义，虽不要求掌握其数学推导，但了解其直观意义、用途和应用条件是必要的，学习时要留心有关解释，并多加思考，这将有助于对公式的理解和正确应用。学习医学统计学还应该多做练习，本书的每一章均配有一定数量的习题，通过做练习，帮助大家学会思考，熟悉概念，学会正确运用统计方法处理实际问题。统计中遇到的计算无非就是加、减、乘、除、平方、开方，再加上查表等，并不复杂。尽管现在有很多统计分析软件包可以省去繁琐的计算，但如果对统计概念理解不透，统计方法选择不当，对计算机打印出的结果亦不会有很深刻的认识。因此，做一些简单的、数据量少的练习是必要的，只有这样才能加深对书本知识的理解，体会出其中滋味。

正确应用统计方法，能帮助我们正确认识客观事物，阐明事物的固有规律，从而把感性认识提高到理性认识。但统计不是万能的，它决不能改变事物的本来面目，把原不存在的规律“创造”出来。有些人在进行试验之前没有充分考虑，收集了一些不准确、不可靠或不全面的资料，希望用统计方法来弥补，这是不可能的，统计只能认识规律而不能“创造”规律。

最后必须注意，统计分析手段需要有正确的医学理论作指导，不能将医学问题归结为纯粹的数量问题，否则会归纳出错误的甚至是荒谬的结论。要知道，医学统计是科学研究的一种工具，它所面对的问题必须来自医学领域；统计学上所得到的结论都具有概率性，它不能证明什么，但可提高你的分辨能力和判断能力，为科学决策提供依据。

1.2 几个基本概念

1.2.1 同质

性质相同的事物称为同质(homogeneity)的，否则称为异质的或间杂(heterogeneity)的。观察单位间的同质性是进行研究的前提，也是统计分析的必备条件，缺乏同质性的观察单位是不能笼统地混在一起进行分析的。如不同年龄组男童的身高不能计算平均数，因为所得结果没有意义。

不同研究中或同一研究中不同观察指标对观察对象的同质性的要求不同，即同质是相对的。例如，男性身高与女性身高有着本质的差别，因此，在考虑身高这一指标时，不能把不同性别的人混在一起，此时，不同性别表示不同质；而在研究白细胞计数这一指标时，因性别对该指标没有影响或影响甚微，故可以把不同性别的人放在一起分析。又如，在某新药的临床试验中，计算有效率的观察病例必须患同一疾病，甚至具有相同的病型、病情、病程等，对同质性的要求是很严格的；而计算不良反应发生率，通常可将不同病种的病例合起来统计，此时对同质性的要求只有一条：按规定服用该新药。

1.2.2 变异

同质的事物之间的差别称为变异。宇宙中的事物千差万别，各不相同，即使是同质事物，就某一观察指标来看，各观察单位(亦称个体)之间也有差别，这种同质事物间的差别称为变异(variation)。例如，研究儿童的生长发育时，同性别、同年龄儿童的身高，有高有矮，各不相同，称为身高的变异。由于观察单位通常是观察个体，故变异亦称个体变异(individual variation)。变异表现在两个方面：其一，个体与个体间的差别；其二，同一个体重复测量值间的差别。变异是宇宙事物的个性反映，在生物学和医学现象中尤为重要。

变异是由于一种或多种不可控因素(已知的和未知的)以不同程度、不同形式作用于生物体的综合表现。如果我们掌握了所有因素对生物体的作用机制,那么,生物体的某指标之观察值就是可预测的了。有些指标的变异原因已被人们认识,例如,染色体决定了新生儿的性别;有些指标的变异原因已被认识一部分,比如,人的身高受遗传和后天营养的影响,但尚有一部分影响因素是未知的;更多的情况下,影响变异的因素是未知的。就每个观察单位而言,其观察指标的变异是不可预测的,或者说是随机的。观察指标用变量(variable)或称随机变量(random variable)来表述。当观察值的个数达到足够多时,其分布将趋于稳定,并最终服从于总体分布。

个体变异现象广泛存在于人体及其他生物体中,是个性的反映。虽然每个个体的变异表现出一定的随机性和不可预测性,但变异并不等于杂乱无章,指标的变异往往是有规律可循的,当所观察的个体数足够多时,观察值的分布将出现一定的规律性,这是总体的反映。从这个意义上讲,变异也是医学研究中必须运用各类统计指标并进行统计分析的缘由,统计学就是探讨变异规律,并运用其规律性进行深入分析的一门学科,可以这么说,没有变异就没有统计学。

1.2.3 总体、个体和样本

总体(population)是根据研究目的所确定的同质观察单位的全体;个体(individual)是构成总体的最基本的观察单位;样本(sample)是从总体中随机抽取的一部分个体;样本中所包含的个体数称为样本含量(sample size)。

例如,调查某地某年正常成年男子的红细胞数,则观察对象是该地某一时间的正常成年男子,全部正常男子构成了一个总体,其同质基础是同一地区,同一年份,同为正常人,同为成年男性;观察单位是该地该年的每一个正常成年男子。如果从中抽取了 20 名,测得其红细胞数,则构成了一个样本含量为 20 的样本。这里的总体只包括(确定的时间、空间范围内)有限个观察单位,称为有限总体(finite population)。有时总体是假想的,如研究某种辅助疗法对肾移植病人生存时间的影响,这里总体的同质基础是同为肾移植病人,同用某种辅助疗法,总体包括设想用该辅助疗法的所有肾移植病人,是没有时间和空间概念的,因而观察单位是无限的,称为无限总体(infinite population)。

医学研究中,很多是无限总体,要直接研究总体的情况是不可能的。即使是有限总体,如果包含的观察单位过多,也要花费大量的人力、物力、财力,有时也是不可能的和不必要的(如检查乙肝疫苗的合格率,不可能将所有的疫苗打开逐一检查),所以实际工作中总是从总体中随机抽取一定含量的样本,目的是根据样本所提供的信息推断总体的特征,这是统计推断的根本内容。

1.2.4 变量的分类

医学科学研究不是对被观察单位(个体)本身感兴趣,而是对个体的某个(些)指标或特征进行观察、测量和分析。这种个体的特征或指标称为变量(variable),变量的取值称为变量值或观察值。

例如,以人为单位,调查某地某年新生儿。“性别”变量的观察结果有男和女;“体重”变量的观察结果有重有轻;“身长”变量的观察结果有长有短;“是否畸形”变量的观察结果有正常、可疑、畸形;“血型”变量的观察结果有 A、B、O、AB 型;“母龄”变量的观察结果亦有大有小;“母

亲曾生胎次”变量的观察结果可取 0,1,2,…;“母亲文化程度”变量的观察结果有文盲、小学、初中、高中、大学等。从上可见,变量的取值可以是定量的,亦可以是定性的。按变量的取值之特性,可将变量分为数值变量和分类变量,不同类型的变量应采用不同的统计分析方法。

1) 数值变量(numerical variable) 或称定量变量,它的取值是定量的,表现为数值大小,一般有度量衡单位,亦称计量资料(上述体重、身长、母龄、胎次均属数值变量),常用第 6、7 章的统计分析方法。

2) 分类变量(categorical variable) 或称定性变量,其取值是定性的,表现为互不相容的类别或属性,有两种情况:

(1) 无序分类(unordered categories) 包括① 二项分类(如上述“性别”变量),表现为互相对立的两种结果;② 多项分类(如上述“血型”变量),表现为互不相容的多类结果。常用第 8 章的统计分析方法。

(2) 有序分类(ordered categories) 各类之间有程度上的差别或等级顺序关系,有“半定量”的意义,亦称等级变量。严格地讲,等级之间只有顺序上的差别而无数值的大小,故等级之间是不能度量的(如上述“母亲文化程度”变量)。常用第 10 章的统计分析方法。

根据分析需要,各类变量间可以相互转化。如上述“体重”变量属数值变量,如按体重小于 2 500g 为低体重儿,大于或等于 2 500g 为正常儿,则“体重”变量可视为二项分类变量。临床上很多检验指标如白蛋白、血小板数、血糖等,可以用具体的数值表示,亦可按临床上的具体表现,将其分为 -,+,++,+++的等级。

1.2.5 统计量与参数

由样本所算出的统计指标称为统计量(statistic)。例如,为了解健康成年男子每升血液中的白细胞数,对一群成年男子进行检验,由所得的一系列数值算出一个算术均数(样本均数)是一个统计量;反映该组数据的变异程度的标准差亦是一统计量。又如,为研究某种畸形的发生率,观察了某年某地出生的所有新生儿,根据该畸形的发生数及新生儿总数求得的畸形发生率是一个统计量。从这些统计量可以估计总体均数、总体标准差、总体率等。这些总体的指标称为参数(parameter)。

总体参数是事物本身固有的、不变的,而统计量则随着试验的不同而不同,但统计量的分布是有规律的,这种规律是统计推断的理论基础。详见第 6 章。

1.2.6 抽样误差

由于总体中每个个体存在着变异,因此从同一总体中随机抽取若干个个体所组成的样本,其统计量如均数、标准差或样本率等,与相应的总体参数不一定恰好相等。这种样本统计量与总体参数间的差别,或不同样本的统计量之间的差别称为抽样误差(sampling error)。

由于生物体的变异总是客观存在的,因而样本的抽样误差是不可避免的,但抽样误差的规律是可以被认识的,因而是可以控制的。“统计推断”就是运用抽样误差的这种规律对总体的某些特征进行估计和推断(详见 6.1.1 节)。

1.2.7 频率与概率

在 n 次随机试验中,事件 A 发生了 m 次,则比值

$$f=\frac{m}{n}=\frac{A\text{ 发生的试验次数}}{\text{试验的总次数}} \tag{1.1}$$

称为事件 A 在这 n 次试验中出现的频率(frequence)。其中 m 称为频数。频率常用小数或百分数表示,显然有:$0\leqslant f\leqslant 1$。医学上通常所说的患病率、病死率、治愈率等都是频率。

如检查某药品的次品率,其结果如下:

抽出样品数 n:	50	100	600	1 500	6 000	9 000	18 000
次 品 数 m:	0	2	7	19	56	93	176
次品率(%)f:	0	2	1.17	1.27	0.93	1.03	0.98

可以看到,抽到次品数的多少具有偶然性,但随着抽取的样品数逐渐增加,次品率 f 将愈来愈接近常数 1%。

实践表明,在重复试验中,事件 A 的频率随着试验次数的不断增加将愈来愈接近一个常数 p,频率的这一特性称为频率的稳定性。

频率的稳定性充分说明随机事件出现的可能,是事物本身固有的一种客观属性,因而是可以被认识和度量的。这个常数 p 就称为事件 A 出现的概率(probability),记作 $P(A)$或 P。这一定义称为概率的统计定义,它是事件 A 发生的可能性大小的一个度量。容易看出,频率为一变量,是样本统计量,而概率为一常数,是总体参数。实践中,当试验次数足够多时,可以近似地将频率作为概率的一个估计。

显然,概率 P 有如下性质:

$$0\leqslant P\leqslant 1 \tag{1.2}$$

常以小数或百分数表示。事件 A 出现的概率愈接近于0,表示 A 出现的可能性愈小;愈接近于1,表示出现的可能性愈大。$P(A)=0$ 表示 A 为不可能事件,即 A 不可能发生;$P(A)=1$ 表示 A 为必然事件,即 A 必然要发生。

按概率的统计定义,为了确定一个随机事件的概率,就得进行大量的重复试验。但有些情况下,可以根据事物本身的性质直接计算某事件的概率。例如,抛掷一枚匀质的硬币,因只有两种可能,且"出现正面"和"出现反面"的机会相等,各占一半,因此,事件 A"出现正面"的概率为 0.5。又如,掷一颗骰子,设骰子是一均匀的六面体,每面分别标有 1 到 6,因掷一次只能出现其中一面,各点出现的可能性相同,所以在一次试验中出现"6 点"的概率为 1/6,而出现"1 点或 6 点"的概率为 2/6。

设某种随机现象具有如下特征:① 所有可能的结果只有有限个,记为 $A_1,A_2,\cdots,A_N$,它们出现的机会均等(等可能性);② 在任一次试验中 $A_1,A_2,\cdots,A_N$ 至少出现其中一种(完备性);③ 在任一次试验中 $A_1,A_2,\cdots,A_N$ 只能出现其中一种(互不相容性)。则在一次试验中 A_i 出现的概率为 $1/N$,出现 A_1 或 $A_2,\cdots$,或 A_M 的概率为 M/N。这一定义称为概率的古典定义。

无论采用何种定义,概率的意义不变,即概率是描述随机事件发生的可能性大小的指标。

1.2.8 小概率事件及小概率原理

医学研究中,将概率小于等于 0.05 或 0.01 的事件称为小概率事件。这种小概率事件虽不是不可能事件,但一般认为小概率事件在一次试验中是不会发生的,这就是小概率原理。小概率原理是统计推断的一条重要原理。详见第 6 章。

2 统计资料的整理与描述

2.1 频数表

对搜集得来的资料(无论是数值变量资料,还是分类变量资料)都要进行整理,使其条理化、系统化,以了解资料的数量特征、分布规律,便于进一步计算统计指标和分析。本节讲述数值变量资料的整理。

2.1.1 频数表的编制

【例 2.1】 某地 120 名 7 岁男童身高(cm)资料如下,试编制频数表。

123.60 121.03 115.42 113.40 124.02 123.41 122.81 125.83 112.33 122.91
124.79 110.12 117.91 126.32 116.55 113.31 114.38 127.22 112.80 120.13
120.62 124.84 117.17 109.85 118.96 116.66 117.44 121.68 118.82 117.63
120.05 119.90 115.24 121.42 125.64 124.24 118.17 120.07 115.12 118.76
116.74 128.35 124.43 115.36 113.59 125.39 120.62 120.10 122.46 120.51
113.26 118.44 122.30 117.36 116.46 121.33 120.88 111.86 117.99 112.65
117.44 124.44 118.69 121.40 118.61 **130.75** 118.31 121.44 117.16 129.65
111.36 115.26 120.78 123.84 123.16 121.23 126.14 118.65 119.19 116.02
115.78 119.01 116.63 120.63 114.30 119.96 116.63 128.41 117.42 123.32
114.09 118.58 116.73 117.11 117.97 **108.13** 126.42 119.66 119.69 118.38
115.16 115.01 119.48 127.58 122.14 122.63 115.57 123.70 123.39 119.59
123.40 119.72 120.60 115.50 123.78 118.41 118.82 114.56 119.45 118.11

频数表的编制方法如下:

1) 找出观察值中的最大值和最小值,并求出极差　本例最大值为 130.75 cm,最小值为 108.13 cm。最大值与最小值之差称为极差(range),记为 R。本例

$$R = 130.75\ \text{cm} - 108.13\ \text{cm} = 22.62\ \text{cm}$$

最大值与最小值反映了观察值的分布范围,极差反映了观察值分布的跨度。

2) 按极差大小决定组段数、组段和组距　频数表一般设 8 ~ 15 个组段,视观察单位数的多少而定。观察数少时,组段可适当少一些,观察数多时,组段可酌情多一些,其原则是要充分反映数据的分布特征。

组距即各组的跨度,是每一组内的范围,常采用等距分组,组距可用下式估计:

$$\text{组距} = \frac{\text{极差}}{\text{组数}}$$

为方便汇总，组距常取整数。

各组段应界限分明，上下衔接，互不交叉，第一组段要包括最小值，最后一组段要包括最大值。每一组段的起点称“下限”，终点称“上限”，为避免交叉，各组段从本组段的“下限”开始（包括下限），到本组段的“上限”为止（不包括上限）。注意，最后一组应同时写出下限和上限。

本例如分 12 组左右，则组距为 22.62/12 = 1.89，取整为 2。第一组段下限为 108，上限为 110，记为“108 ~ ”，包括最小值；第二组段下限为 110，上限为 112，记为“110 ~ ”；最后一组段下限为 130，上限为 132，记为“130 ~ 132”，包括最大值。见表 2.1 第(1)栏。

3) 列表划记　计算各组段包含的观察单位个数，即频数。各组频数之和应等于总观察数。见表 2.1 第(2)、(3)栏。

表 2.1　120 名 7 岁男童身高(cm)的划记和频数

身高组段 (1)	划　记 (2)	频数 (3)
108 ~	丅	2
110 ~	下	3
112 ~	正丅	7
114 ~	正正𠂇	14
116 ~	正正正𠂇	19
118 ~	正正正正𠂇	24
120 ~	正正正下	18
122 ~	正正正	15
124 ~	正𠂇	9
126 ~	正	5
128 ~	下	3
130 ~ 132	一	1
合　计		120

2.1.2　频数分布的图示

以身高为横轴，以频数为纵轴，每一组段画一直条，直条的面积与该组频数成正比，如图 2.1 所示，称为直方图(histogram)。

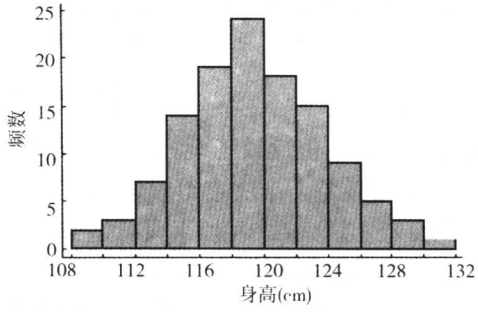

图 2.1　某地 120 名 7 岁男童身高(cm)的频数分布

2.1.3　频数分布的分析

频数表或频数分布图作为陈述资料的一种基本形式常见于文献、科研报告、工作总结和统计报表中，被称为加工过的资料。对频数表的分析，主要在于以下几个方面：

1) 有无可疑值　通过对频数分布的分析，发

现某些特大或特小的离群值(outlier)、可疑值。例如,有时在频数表的两端,出现连续几个组段的频数为0后,又出现一些特大或特小的值,使人怀疑这些数据是否准确,需进一步核查,如有错应予纠正。

2) 分布的类型　频数分布可分为对称分布和偏态分布两种类型。所谓对称,是指观察值向中央部分集中,以中等数据居多,左右两侧分布大体对称。所谓偏态分布,是指观察值偏离中央,尾部偏向数轴正侧,称正偏态;反之,尾部偏向数轴负侧,称负偏态。如食物中毒引起腹泻的潜伏期,一般在几个小时之内,但也有个别拖到十几个小时的,其分布为正偏态;又如,某些慢性病患者以老年人为主,则其年龄分布偏向于年龄大的一侧,为负偏态分布。不同类型的分布,应采用不同的统计分析方法,如例2.1资料的分布属对称分布。

3) 分布特征　从频数表还可看到分布的两个重要特征,即集中趋势(central tendency)和离散趋势(tendency of dispersion)。集中趋势在表2.1资料中表现为120名男童的身高大多集中在"118～"cm左右;但120个数据仍参差不齐,从最小的108.13 cm到最大的130.75 cm,且由中间向两侧逐渐减少,数据的这种分布特征体现了离散趋势。

2.2　集中趋势的描述

平均数(average)反映一组观察值的集中趋势、中心位置或平均水平。它是该组数据的代表,能对一群同类事物或现象的数量特征作出概括的说明,是统计学中应用最广泛、最重要的一个指标体系。常用的平均数有均数、几何均数和中位数,分述如下。

2.2.1　均数

均数(mean)是算术均数(arithmetic mean)的简称,习惯上用希腊字母 μ(读作 mu)表示总体均数,用 $\bar{X}$(读作 X bar)表示样本均数。均数反映一组观察值在数量上的平均水平,最适用于单峰对称分布资料的平均水平的描述。

1) 未分组资料的均数计算法　将所有观察值 $X_1, X_2, \cdots, X_n$ 直接相加,再除以总观察数 n,即得均数。以公式表示为

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum X_i}{n} \tag{2.1}$$

式中 $\sum$(读作 sigma)是求和的符号,$\sum X_i$ 表示对所有观察值 $X_1, X_2, \cdots, X_n$ 求和。

【例2.2】 10名20岁女青年血清总蛋白含量(g/L)如下,试求其均数。

74.3,　75.6,　78.8,　67.2,　70.4,　77.6,　81.6,　67.3,　70.3,　71.2

$$\bar{X} = \frac{74.3 + 75.6 + 78.8 + 67.2 + 70.4 + 77.6 + 81.6 + 67.3 + 70.3 + 71.2}{10} = 73.43(\text{g/L})$$

2) 分组资料的均数计算法　有时我们面对的资料不是原始数据,而是经过加工整理后的分组资料,这时可用加权法求均数,有

$$\bar{X} = \frac{f_1 X_1 + f_2 X_2 + \cdots + f_k X_k}{\sum f_i} = \frac{\sum f_i X_i}{\sum f_i} \tag{2.2}$$

式中 f_i 是第 i 组的频数，$\sum f_i$ 表示各组频数之和，即总观察数 n；X_i 是第 i 组的组中值，即该组的(下限 + 上限)/2。由于只知道有 f_i 个观察值属于该组段，而不知道具体的数值，故该组的 f_i 个观察值均以组中值代替。显然，从频数表求出的均数与直接从原始数据所求得的均数稍有出入，在有原始数据的情况应尽量用原始数据直接计算。

【例 2.3】 求表 2.1 中资料的均数。

用公式(2.2)求频数表资料的均数为

$$\bar{X} = \frac{\sum f_i X_i}{n} = \frac{2 \times 109 + 3 \times 111 + 7 \times 113 + \cdots + 3 \times 129 + 1 \times 131}{120} = 119.43(\text{cm})$$

直接求原始资料的均数为 119.41(cm)，两者稍有出入，但在单峰对称分布时近似程度甚好。

2.2.2 几何均数

有些医学资料，如抗体的滴度、细菌计数等，其频数分布呈明显偏态，各观察值之间呈倍数变化(等比关系)，算术均数对这类资料集中趋势的代表性较差，这时宜用几何均数(geometric mean)反映其平均增(减)倍数。几何均数一般用 G 表示，适用于各变量值之间成倍数关系，但作对数变换后指标成单峰对称分布的资料。

1) 未分组资料的几何均数计算法　将 n 个观察值 $X_1, X_2, \cdots, X_n$ 直接相乘，再开 n 次方，即为几何均数。以公式表示为

$$G = \sqrt[n]{X_1 X_2 \cdots X_n} \tag{2.3}$$

当各观察值甚小(接近于 0)或过大，或当 n 较大时，连乘运算常使计算器(机)内存溢出，因而无法运算，这时可借助于对数变换来计算。即先求各观察值的对数值之算术均数，再用反对数变换得其几何均数。以公式表示为

$$G = \lg^{-1}\left[\frac{\lg X_1 + \lg X_2 + \cdots + \lg X_n}{n}\right] = \lg^{-1}\left[\frac{\sum \lg X_i}{n}\right] \tag{2.4}$$

【例 2.4】 5 人的血清抗体滴度为 1:2,1:4,1:8,1:16,1:32，求平均滴度。

由于数据间呈倍数关系，以用几何均数为宜。先求滴度倒数的平均，有

$$G = \sqrt[5]{2 \times 4 \times 8 \times 16 \times 32} = 8$$

或

$$\lg G = \frac{\lg 2 + \lg 4 + \lg 8 + \lg 16 + \lg 32}{5} = 0.903, \quad G = \lg^{-1} 0.903 = 8$$

故平均滴度为 1:8。

2) 分组资料的几何均数计算法　以公式表示为

$$G = \lg^{-1}\left[\frac{f_1 \lg X_1 + f_2 \lg X_2 + \cdots + f_k \lg X_k}{\sum f_i}\right] = \lg^{-1}\left[\frac{\sum f_i \lg X_i}{n}\right] \tag{2.5}$$

【例 2.5】 某地 55 人接种疫苗后抗体滴度见表 2.2 第(1)、(2)栏，求平均滴度。